Big Data Midterm

Question 1

from mrjob.job import MRJob

from mrjob.step import MRStep

import re

regex = re.compile(r"(^\d+\w+);(.\*\w+\s\*\w\*.\*);(\w+);(\w+.);(\d+.\d+);.(\d{4})")

class MR1(MRJob):

    def steps(self):

        return [MRStep(mapper=self.mapper,

        combiner=self.combiner\_count\_words,

        reducer=self.reducer\_count\_words)]

    def mapper(self,\_, line):

        data = regex.findall(line)

        for item in data:

            id = item[0]

            city = item[1]

            state = item[2]

            payment\_type = item[3]

            payment\_value = item[4]

            year = item[5]

            if year == '2017' and payment\_type =='credit\_card':

                yield (payment\_type, 1)

    def combiner\_count\_words(self, word, counts):

        # optimization: sum the words we've seen so far

        yield word, sum(counts)

    def reducer\_count\_words(self, word, counts):

        # send all (num\_occurrences, word) pairs to the same reducer.

        # num\_occurrences is so we can easily use Python's max() function.

        yield (word,sum(counts))

if \_\_name\_\_ == '\_\_main\_\_':

    MR1.run()

a) 759

b) 330

c)

Question 2

from mrjob.job import MRJob

from mrjob.step import MRStep

import re

class MR1(MRJob):

    def steps(self):

        return [MRStep(mapper=self.mapper,

        combiner=self.combiner\_count\_words,

        reducer=self.reducer\_count\_words)]

    def mapper(self,\_, line):

        data = line.split()

        for item in data:

            if item == "Bennet":

                yield (item, 1)

    def combiner\_count\_words(self, word, counts):

        # optimization: sum the words we've seen so far

        yield word, sum(counts)

    def reducer\_count\_words(self, word, counts):

        # send all (num\_occurrences, word) pairs to the same reducer.

        # num\_occurrences is so we can easily use Python's max() function.

        yield (word,sum(counts))

if \_\_name\_\_ == '\_\_main\_\_':

    MR1.run()

1. Bennet has appeared 14 times

Question 3

SELECT COUNT(\*), payment\_type FROM question1\_olist\_v1.csv

GROUP BY pyament\_type

Question 4

Question 7

1. \$\w+
2. #\w+